12-7-2020

Can Statcast variables explain the variation in weighted runs created plus?

Ryan Kupiec
DePauw University, ryankupiec_2021@depauw.edu

Can Statcast variables explain the variation in weighted runs created plus?

Ryan Kupiec

DePauw University

12/7/2020

Abstract

The release of Statcast data in 2015 was revolutionary for data analysis in the game of baseball. Many analysts have begun using this data regularly, but none have used it exclusively. Often older, less reliable statistics (on-base percentage) are still used in favor of the newer statistics (weighted runs created plus). In this paper, we attempt to explain the variation in weighted runs created plus (wRC+) using Statcast variables such as exit velocity and launch angle. We find that exit velocity along with other Statcast variables, can explain as much as 70% of the variation in wRC+. Launch angle can significantly explain the variation in wRC+ but did not provide enough to the model to warrant keeping.

Keywords: Sabermetrics, wRC+, Statistics, Data Analysis**,** Exit Velocity

Introduction

In the last two decades, baseball has seen a revolution, unlike any we have seen in sports. Interestingly enough, this revolution is not a physical performance or tangible item; but rather a statistical revolution. This radical and unprecedented change has transformed the way players are evaluated, the way teams approach the game, and the way players play the game. Michael Lewis' *Moneyball* in 2003 flipped the game of baseball on its head, calling into question every previous notion of how baseball should be played. Some teams have been quicker to adapt to the new landscape of Major League Baseball (MLB) while others have been more hesitant; however, the overwhelming statistical effect is undeniable.

To the dismay of baseball traditionalists, it appears as though this statistical revolution is here to stay. The term "sabermetrics" has been attached to these advanced baseball metrics and has become an integral part of the game. The sabermetrics movement took another step forward in 2015 when the MLB installed cameras/sensors called Statcasts in all 30 parks. Statcast further

built upon the cameras already installed by Major League Baseball Advanced Media: "Using a combination of radar and video, Statcast now provides over 17 petabytes of data each season on the movement of every player and ball during all Major League games."[18] With new data comes new statistics. More importantly, with those recent statistics comes the possibility of new solutions to questions never before asked. Statcast has opened the door to an immense amount of questions.

Moneyball was groundbreaking because it steered attention from poor statistics of player evaluation to more accurate statistics for player evaluation. However, the statistics available at the time of *Moneyball* were all results-oriented statistics, meaning they counted after the play occurred. Statcast, in comparison, delivers statistics that can measure athleticism during a play. These new statistics offer more ways to evaluate players and even replace the statistics that we initially believed to be the most accurate. To showcase the power of this new Statcast data, this paper will attempt to see how much variation in weighted runs created plus (wRC+) can be explained using a number of different Statcast variables for 2019 MLB data. The Statcast variables we will be using for the analysis include: exit velocity (EV), barrel percentage (Barr%), launch angle (LA), walk percentage (BB%), age, speed score (Spd), opposite-field percentage (Oppo%), pull field percentage (Pull%), line drive percentage (LD%), fly ball percentage (FB%), and Strikeout percentage (K%).

Literature Review

Baseball terminology is a continually evolving landscape. Every year, there seems to be new terms used to measure players to the extent that the casual fan can get lost in the myriad of definitions used in regular everyday broadcasts. With this in mind, we will evaluate key terminology for this paper's research.

The independent variable of interest is weighted runs created plus (wRC+). The wRC+ is a rate statistic that attempts to fully encompass a player's offensive value. The measure, wRC+, weights each offensive outcome differently and controls for park and league effects.[25] This wRC+ is scaled so that 100 is average every year, and 1 point above or below 100 is equal to one percentage point better or worse than league average.[25] This allows wRC+ to be compared across leagues and years. The formula for wRC+ is:

$$wRC+ = \frac{(((wRAA/PA + League\ R/PA) + (League\ R/PA - Park\ Factor * League\ R/PA))}{(AL\ or\ NL\ wRC/PA\ excluding\ pitchers))} * 100$$

The different ranges of wRC+ production are found in table 1.

Exit velocity (EV) is a Statcast metric that "measures the speed of the baseball as it comes off the bat, immediately after a batter makes contact."[12] Exit velocity is tracked for all batted ball events: hits, outs, and errors.[12] Hitting the ball hard does not guarantee getting on base; however, more often than not, the harder you hit the ball, the better your chances are of a positive outcome.[3] Exit velocity is skill-based (in control of the hitter) as the hitter controls how hard they hit the ball. Another advantage of EV is that it stabilizes much faster than other traditional statistics.[3] This means that it reaches a correlation of 0.7 with itself (in as quick as 40 balls in play) much faster than the on-base percentage (350 plate appearances).[7]

Ratings	wRC	wRC+
Excellent	105	160
Great	90	140
Above Average	75	115
Average	65	100
Below Average	60	80
Poor	50	75
Awful	40	60

Table1: Table of the values of wRC+ and their corresponding rating. Source: https://library.fangraphs.com/offense/wrc/

Launch Angle (LA) is the vertical angle at which the ball leaves a player's bat after being struck.[13] A ground ball is less than 10 degrees, a line drive is 10-25 degrees, a fly ball is 25-50 degrees, and a pop up is greater than 50 degrees.[13] Launch angle can give us insight into what type of hitter is up at-bat. Without an ideal launch angle, it is impossible to get a hit as you will either hit the ball straight into the air or straight into the ground. The ideal launch angle is considered to be 25-35 degrees by analysts.[22] However, an important note is that you must also have a high exit velocity to take that launch angle anywhere, which leads to the next variable.

A barrel is essentially a hit type classification. In order for a batted ball to be classified as a barrel, the ball must have a combination of EV and LA that equate to at least a minimum .500 batting average and 1.500 slugging percentage since Statcast was implemented in 2015.[15] Batted balls that are struck with at least a 98 mph EV and between 26-30 degrees LA are classified as a barrel.[15] For every mile per hour over 98, the range of launch angles expands.[15] For this analysis, we will be using barrel percentage (Barr%). We will employ barrel percentage because we are looking at predicting players' offensive impacts over an entire season, so it makes more sense to look at the percentage of the players' batted balls that resulted in the coveted barrel classification. To use only barrels gives an advantage to players who had more batted balls or more plate appearances, which would have opened up the opportunity for more barrel opportunities.

Walk percentage (BB%) is the frequency with which the batter has walked.[30] The BB% is calculated by dividing walks by the total number of plate appearances. Strikeout percentage (K%) is another frequency statistic that will be added to the model. The K% is the frequency with which a batter has struck out. K% is calculated by the number of strikeouts

divided by the number of plate appearances. It is crucial to have K% and BB% as they give insight into the style of the hitter. Players with high K% and low BB% are typically hitters that swing more aggressively which results in more strikeouts than walks. High BB% and low K% suggest more of a conservative approach at the plate.

Speed Score (Spd) is a statistic that was created by sabermetric legend Bill James. Spd is a statistic that rates a player based on their speed and baserunning abilities.[24] Due to the belief that faster players can turn some outs into hits, it is essential that a speed variable be added to the model. Inversely, slower players could potentially be thrown out more often than they should. Spd is on a scale from zero to ten, with four and a half being the average, two being poor, and seven being excellent.[24]

Pull (pull%), center (cent%), and opposite (oppo%) percentages are the percentages of the time a player hits the ball to a given part of the field. The field is separated into three sections. Depending on the side of the plate you hit from (left or right), these thirds have different names. Using a right-handed hitter as an example: hitting the ball to the left side is pulling the ball, hitting up the middle is center, and hitting the ball to the right side is the opposite field. This is important to add because in 2018, "32.7 percent of fly balls to a batter's pull side went for home runs, compared to 8.1 percent of fly balls to center field and 3.8 percent to the opposite field. Batters across the league had a .429 average, and 1.514 slugging percentage on fly balls hit to the pull side, and a 0.135 average and .324 slugging mark on balls hit to the opposite field."[21] It is often believed that power hitters pull the baseball and these numbers support this belief. Therefore, we believe these are important to look at because better hitters are more likely to pull the baseball and, when they pull the ball, land more hits improving their wRC+.

Fly ball (FB%), line drive (LD%), and ground ball (GB%) percentages are the percentage of the time that a player hits a ball classified as a fly ball, line drive, or ground ball. These batted ball statistics are calculated by an algorithm used by Baseball Info Solutions.[23] Therefore, we cannot gain the criteria for the ball to be classified as any of these three events. These variables are important to add as line drives tend to signal a player making good contact with the baseball. In recent years, fly balls have been sought after with the belief that getting the ball in the air more often could increase home runs. So it will be essential to look at the effects each has on wRC+.

We will use FB%, GB%, LD%, and pull%, cent%, and oppo%, as proportions. Typically, both of these sets of batted ball statistics are percentages. Still, for this research, we will be using them as proportions. We look at them as proportions because both sets add up to one when you add all three. Furthermore, we will exclude one of each set from the model as adding all three would create high variable inflation factors that indicate multicollinearity.

Age is the final independent variable in the model. The peak age of MLB players is still debated among the baseball community. Nonetheless, we do know that it is more difficult to calculate the effects of age than you would think.[6] However, we do know that production slips with age. Therefore, it is crucial to have an age variable in the model to account for players growing older.

Past Research

Baseball research is a unique field in that most of the research is posted online as articles for the common fan to read. Typically if one is good enough to produce studies, they will be scooped up by MLB teams as has been seen with many writers for one of the most popular

baseball article platforms: Fangraphs. However, Statcast is such a hot topic among the baseball community that there is a plethora of previous research to explore.

In an article in 2015, Rob Arthur explored the relationship between exit velocity (EV) and on-base plus slugging percentage (OPS). Arthur found that for every additional mile per hour of batted ball velocity, it equated to an increase of 18 points in OPS.[3] In sum: players who hit the ball harder tend to get on base more.[3] Arthur found this with an R-squared value of .1475, which he classified as a significant relationship. A graph with the model can be found in figure 1. Arthur also explained in the article that correlation does not imply causation as many players hit the ball hard but do not have a starting level OPS.

Building off of Arthur and other sources, Nicholas Taylor wrote a Forecasting Batter Performance thesis using Statcast data in 2017. Taylor attempted to test if a player's average exit velocity could significantly explain the variation in a player's batting average on balls in play

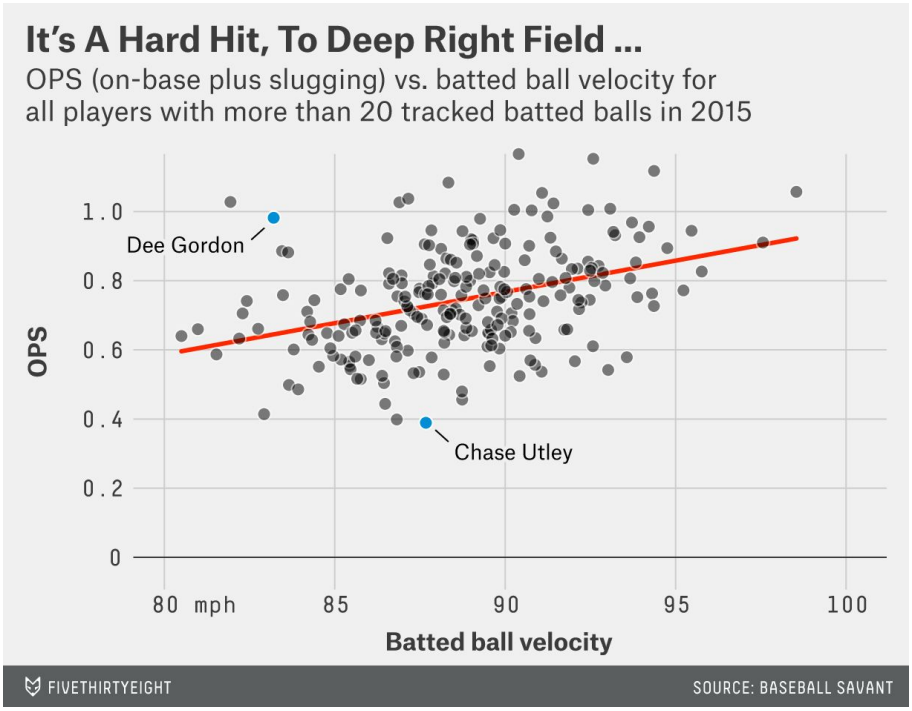


Figure 1: a fitted line plot predicting OPS using Exit velocity for early season MLB 2015 data

(BABIP) and slugging percentage.[27] Taylor's model also used other Statcast variables. Taylor found that the model can explain half of the variation in a player's slugging percentage, with EV being one of the most significant variables.[27] Taylor also found that each additional mile per hour in exit velocity accounted for almost 10 additional points to slugging percentage.[27] The model did not produce as well with BABIP as it only accounted for about forty percent of the variation in BABIP.

More similar to this paper, Gavin Sanford examined the Statcast variables with the most significant effect on wRC+ on 2017 MLB data. Sanford found that barrels per plate appearance, walk rate, strikeout rate, line-drive percentage, sprint speed, and the number of times you hit the ball 95 mph+ are all statistically significant in determining wRC+.[20] Sanford also looked at players who under and over performed their expected wRC+ based on the created model. Sanford found that barrels could be a better predictor than exit velocity or launch angle.[20]

Previous research performed using Statcast data indicates that many variables such as exit velocity are positively correlated and good indicators of advanced metrics like slugging percentage and wRC+. The research confirms that hitting the ball harder increases your chances of getting on base and hitting extra-base hits. The study also hints at the versatility of exit velocity as it explains variation in BABIP, slugging percentage, OPS, and wRC+. However, the research also indicates that there could be more advanced metrics coming, such as barrels that could be better at explaining the variation in baseball metrics.

Methodology

The purpose of this research is to see how much variation in wRC+ can be explained using exit velocity (EV) and launch angle (LA) along with many other Statcast variables. To test this, we will employ an ordinary least squares regression model. The model will use wRC+ as

the dependent variable and Statcast variables as the independent variables. The Statcast variables representing the initial independent variables include: exit velocity (EV), barrel percentage (Barr%), launch angle (LA), walk percentage (BB%), age, speed score (Spd), opposite-field percentage (Oppo%), pull field percentage (Pull%), line drive percentage (LD%), and strikeout percentage (K%). We will first generate correlation plots between the independent variables and wRC+. This will be done to see how significant each variable is to the model. From there, we will determine which variables are significant enough to keep in the model. Those not significant enough will be dropped from the model. The full, backward, and forward stepwise regression will then be employed to ensure that we have the most critical variables for the model. Models will be compared using R-squared, adjusted R-squared, Mallows CP, AIC, and root mean square error (RMSE). After selecting the model, we will use the variance inflation factor (VIF) to ensure no multicollinearity within the variables. We will consider the cutoff for the VIF to be 5, although VIF=10 confirms severe multicollinearity. From here, we will run the model and run the residuals analysis to see how well the model performed.

The data used for the models will be taken from the python package "pybaseball". Pybaseball is a Python package that scrapes websites such as Baseball Reference, Baseball Savant, and Fangraphs. We will use this package to pull data from MLB's Statcast database into Python. Then we will zip it into a CSV file before importing it into R. All data will be taken from the 2019 MLB Statcast database. An important caveat is that we will only be using players with 100 or more batted ball events to have an adequate sample size to gauge the players.

Results

After deleting all samples from the data set with less than one hundred batted ball events, we are left with a sample size of 406 players from the 2019 MLB season. We will first examine a

correlation plot between each of the independent variables and the wRC+. This will allow us to compare the relationship between the dependent and independent variables.

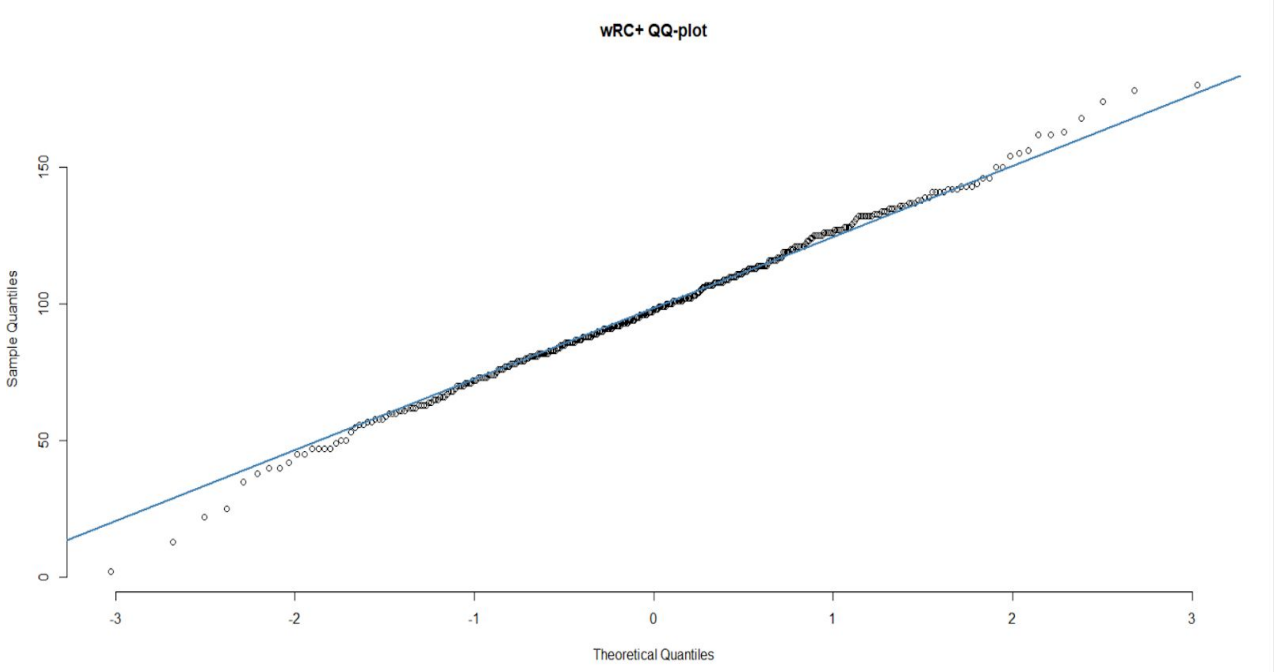
	wRC+	Age	Spd	Pull%	Oppo%	LD%	FB%	EV	LA	Barrel%	BB%	K%
wRC+	1											
Age	-0.071	1										
Spd	0.068	-0.245	1									
Pull%	0.101	0.057	-0.094	1								
Oppo%	-0.087	-0.017	0.073	-0.819	1							
LD%	0.25	0.013	0.03	-0.212	0.226	1						
FB%	0.188	-0.019	-0.059	0.539	-0.452	-0.227	1					
EV	0.549	-0.031	-0.107	0.087	-0.134	-0.064	0.193	1				
LA	0.175	0.006	-0.049	0.504	-0.403	0.041	0.923	0.08	1			
Barrel%	0.577	-0.101	-0.127	0.282	-0.257	-0.028	0.396	0.746	0.309	1		
BB%	0.406	0.09	-0.116	0.175	-0.135	0.099	0.272	0.329	0.25	0.377	1	
K%	-0.185	-0.174	0.011	0.07	-0.033	-0.069	0.208	0.198	0.174	0.424	0.137	1

Table 2: Correlation matrix between wRC., Age, Spd, Pull., Oppo., LD., FB., EV, LA, Barrel., BB., K.

According to the graph, we observe that barrel percentage is the highest correlated variable with wRC+ (r=.577). This is not surprising as a "barrel" is considered a near optimal swing on the baseball and therefore, should turn into offensive production. The next highest correlated variable is exit velocity (r=.549). This bodes well for our hypothesis that exit velocity would be one of the top variables that influence wRC+. Squaring the correlation coefficient yields an r-squared value of .301. This means that 30.1% of the variation in wRC+ can be

explained by exit velocity alone. While that may not seem significant, explaining +30% of the variation of such an integral statistic as wRC+ speaks to the power of average exit velocity.

The next step is to use analysis of variance (ANOVA) to check the performance of the ordinary least squares regression and evaluate the variation in wRC+ that the independent variables can explain. Before we can run ANOVA, however, we must first check the normality assumption of the response variable. We will look at a histogram to ensure wRC+ is normally distributed and also create a quantile-quantile (Q-Q) plot. The variable, wRC+ appears to be normally distributed according to the histogram and passes the normality assumption. wRC+ also does not appear to show any signs of an s-shape in the QQ-plot. We can now move onto ANOVA for further testing.



Our initial model produced an r-squared value of 0.6938. This means that almost 70% of the variation in wRC+ can be explained by our independent variables. Our adjusted r-squared is 0.6852. This is not a significant decrease in our r-squared value. Adjusted r-squared penalizes the

model for adding variables that do not help the model. A massive difference between r-squared and adjusted r-squared would indicate that the model contains unnecessary variables.

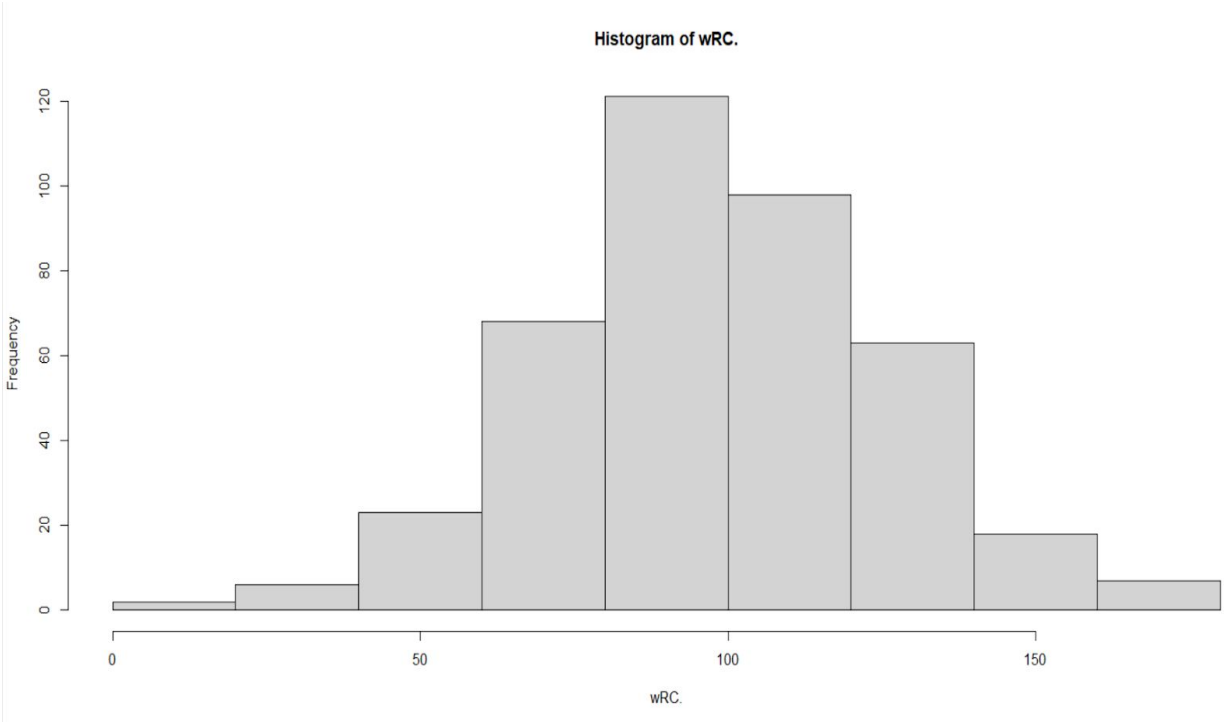


Figure 2: Histogram of the distribution of wRC+ samples.

In table 3, the column titled Pr(>F) measures the probability that the acquired F-value could have happened had none of the independent variables had an effect on wRC+. We are still testing at an alpha level=0.05. Considering many of the variables have a p-value<0.05, it appears as though many of our independent variables have significant effects on wRC+. According to the initial p-values, pull% (p-value< 0.0001), LD% (p-value < 0.0001), FB% (p-value < 0.0001), EV (p-value < 0.0001), LA (p-value < 0.001), Barrel% (p-value < 0.0001), BB% (p-value < 0.0001), and K% (p-value < 0.0001) are all extremely significant with p-values below .001.

Our two variables of most interest (exit velocity and launch angle) performed well given the initial ANOVA test. The larger the F-value, the less likely that the variation in wRC+ caused by the independent variables happened by chance. With exit velocity having the largest F-value (372.711) and a low p-value (p-value < 0.0001), the ANOVA test is affirming the correlation

matrix suggestion that EV explains a good amount of the variation in wRC+. Launch angle also appears to be significant (p-value < 0.001 and F-value=11.580).

In contrast, speed score (p-value=0.0618) and opposite field percentage (p-value=0.8986) appear to be insignificant based on their p-values. Judging by the simultaneous low sum of squares, it appears that the opposite field percentage does not add much to the model. This is

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Age	1	1573.6159196	1573.6159196	6.5707244	0.0107373
Spd	1	840.0371362	840.0371362	3.5076237	0.0618273
Pull%	1	3718.6764944	3718.6764944	15.527549	9.62e-05
Oppo%	1	3.8921522	3.8921522	0.0162519	0.8986232
LD%	1	24258.3365168	24258.3365168	101.2920888	< 2e-16
FB%	1	11755.3827412	11755.3827412	49.0852813	< 2e-16
EV	1	89260.3702605	89260.3702605	372.7118447	< 2e-16
LA	1	2773.2525328	2773.2525328	11.5798765	0.000735
Barrel%	1	11468.8376025	11468.8376025	47.8887955	< 2e-16
BB%	1	7293.801203	7293.801203	30.4556893	< 2e-16
K%	1	60852.4890164	60852.4890164	254.0930916	< 2e-16
Residuals	394	94358.6483259	239.4889551		
Model Summary					
R^2	Adjusted R^2				
.6938	.6852				

Table 3: ANOVA output of the full regression model predicting the variation in wRC+.

expected, as in accordance with past research, players do not fare well when hitting the ball to the opposite field. Speed score appears to be more significant to the model with a much larger sum of squares and passing at the 0.10 alpha level. However, the speed score still does not pass

at our established alpha level of 0.05 for this test. A possible explanation that speed score is not as significant as the other variables is that speed score is normally distributed. With many players near the average speed score, speed probably does not influence enough wRC+ outside of the average. We will run further tests to determine whether to keep or drop some variables.

Before we continue with our model analysis, it is essential to check our independent variables for multicollinearity. Multicollinearity is when there are near-linear dependencies among the regressors.[1] This can cause certain variables to get too much or too little credit in the model essentially causing the inferences to be misleading. To check for this, we will look at the variance inflation factors (VIFs). Any VIFs between 5 and 10 indicate a high correlation.[1] Any VIFs above 10 suggest the associated regression coefficients are poorly estimated because of multicollinearity.[2]

Examining the VIFs, we notice there are two variables with high VIFs: FB% and LA. We saw with the correlation matrix that FB% and LA were highly positively correlated (r=0.923). It makes sense that LA and FB% are associated, given that a launch angle essentially dictates whether you hit a fly ball or not. When multicollinearity is present in your model, there are several possible solutions. One possible solution would be to collect more data and exclude the correlated variables. However, collecting more data would most likely not alleviate this issue, given that the launch angle's relationship with fly balls will remain. Another option is the model respecification. There does not seem to be a possible way to transform these two variables, however. The simplest way to handle our current issue of multicollinearity is variable elimination. You must be careful with variable elimination as the variable being dropped cannot

[1] The book
[2] The book

have significant explanatory power on wRC+, or we risk damaging the model's predictive power.[3]

Variables	VIFs
Age	1.1346
Spd	1.0975
Pull%	3.591
Oppo%	3.124
LD%	2.1955
FB%	15.4688
EV	2.6447
LA	14.081
Barrel%	3.4467
BB%	1.2735
K%	1.3306

Table 4: Variance inflation factor values for each variable in the current model.

Both FB% and LA were significant when looking at the p-value of the F-test. However, we will drop FB% from the model. From a relevant standpoint, the launch angle makes more sense. The scope of this paper is to evaluate the variation in wRC+ that newer Statcast variables can explain, more specifically launch angle and exit velocity. The launch angle is one of the hottest topics in baseball, creating many arguments between baseball traditionalists and sabermetricians. Fly ball percentage is useful but a near afterthought in comparison to the launch angle.

A critical clarification must be made with fly ball percentage being dropped. We were originally using line drive percentage, fly ball percentage, and ground ball percentage as proportions because together, they add up to one. With fly ball percentage being dropped and the line drive percentage being kept, line drive percentage is no longer a proportion of batted balls

[3] The book p 304

with fly balls and ground balls. The line drive percentage is now a proportion with all batted balls. This means that the line drive proportion is now the proportion of batted balls that were hit for a line drive. The inverse of the line drive percentage is a batted ball event that did not end in a line drive. Now we will reevaluate the VIFs for the variables of the new model.

As we can see in table 5, removing fly ball percentage eliminated the multicollinearity that was present in the model. The VIFs are now all under 5, with the highest one being pull percentage with a VIF of 3.59. This suggests that there is no presence of multicollinearity in our model. Now that we have handled this assumption, we will move into the further assessment of the model through the process of stepwise regression.

Variables	VIFs
Age	1.1306
Spd	1.0967
Pull%	3.5909
Oppo%	3.124
LD%	1.1297
LA	1.5083
EV	2.613
Barrel%	3.3769
BB%	1.2538
K%	1.328

Table 5: Variance inflation factor values for each variable in the updated model.

When looking at stepwise regression, there are three broad stepwise procedures (forward selection, backward elimination, and stepwise regression) that are employed. We will run all of them to see the suggested models and then compare many different metrics to see which model performs the best. We will start with stepwise regression. Our goal will be to match the top

performers in all three procedures to find the highest performing model. To measure the top-performing models, we will be looking at R-squared, adjusted R-squared, mallow's Cp, and AIC.

The table (Table:5) has the top performance at each number of parameters until you reach the full model for stepwise regression. When considering every metric available from stepwise regression, it appears as though n=6 or n=7 is our top model choice. Any model with less than six variables has a large mallow's Cp (we are looking for a mallow's Cp close to the number of variables). In contrast, the models with greater than seven variables do not appear to be substantially increased in R-squared or adjusted R-square. With this in mind, we will turn to backward elimination.

The results for backward elimination can be seen in table 7. Backward elimination suggests a model with 8 variables. The variables recommended include Age, Spd, Oppo%, LD%, EV, Barrel%, BB%, and K%. This agrees with the 8 variable model that was suggested by the stepwise regression. We will check these suggestions with the forward selection before making a final decision.

The final output for the forward selection can be found in Table 8. We will use forward selection to see if there are any other models to evaluate before considering the suggestions from backward elimination and stepwise regression. The forward-selection will prioritize p-value as the criteria for evaluating variables. Forward selection suggested the same model as backward elimination and one of the top suggestions from stepwise regression. The forward selection states that opposite-field percentage is the most significant variable with an adjusted R-squared of 0.6840. This is interesting because every other measure we have seen has listed the

N	Predictors	R-Square Adj.	R-Square	Mallow's Cp
1	Barrel.	0.333260813	0.331610469	450.500836
2	Barrel. K.	0.558696062	0.556505968	164.256584
3	LD. Barrel. K.	0.614755127	0.611880165	94.578782
4	Spd LD. Barrel. K.	0.643619775	0.640064860	59.672092
5	Spd LD. Barrel. BB. K.	0.675403332	0.671345874	21.033248
6	Spd LD. EV Barrel. BB. K.	0.683804495	0.679049676	12.291419
7	Age Spd LD. EV Barrel. BB. K.	0.688247160	0.682764070	8.610974
8	Age Spd Oppo. LD. EV Barrel. BB. K.	0.690251531	0.684009748	8.048161
9	Age Spd Oppo. LD. LA EV Barrel. BB. K.	0.690972344	0.683948988	9.126521
10	Age Spd Pull. Oppo. LD. LA EV Barrel. BB. K.	0.691071295	0.683250316	11.000000

Table 6: Stepwise regression of each model at each level of n.

Variable	Df	Sum of Sq	RSS	AIC
<none>			95451	2234.8
Oppo.	1	618	96069	2235.4
Age	1	1325	96776	2238.4
EV	1	2310	97761	2242.5
Spd	1	7878	103329	2265.0
BB.	1	9935	105386	2273.0
LD.	1	12002	107454	2280.9
Barrel.	1	42109	137560	2381.1
K.	1	61571	157022	2434.9

Table 7: Backwards elimination of the model.

Step	Variable Entered	R-Square	Adjusted R-Square	C(p)	AIC	RMSE
1	Barrel.	0.3333	0.3316	450.5008	3686.2005	22.5514
2	K.	0.5587	0.5565	164.2566	3520.6585	18.3697
3	LD.	0.6148	0.6119	94.5788	3467.5015	17.1847
4	Spd	0.6436	0.6401	59.6721	3437.8818	16.5490
5	BB.	0.6754	0.6713	21.0332	3401.9554	15.8135
6	EV	0.6838	0.6790	12.2914	3393.3090	15.6271
7	Age	0.6882	0.6828	8.6110	3389.5641	15.5364
8	Oppo.	0.6903	0.6840	8.0482	3388.9454	15.5059

Table 8: Forward selection of the model.

importance of variables the opposite way that forward selection has. With this in mind, the forward selection will carry less weight in our decision in comparison to backward elimination and stepwise regression.

With all three models in mind, we must now evaluate the best regressors for the model. Even though all three stepwise-type procedures suggest the 8 variable model, I believe that the best model for wRC+ is the 6 variable model. Looking back at the stepwise regression, the difference between the 6 and 8 variable models is minimal. The change in models results in a change in adjusted r-squared of less than .01. Even less of a difference for the r-squared value. By switching to the 6 variable model, we do accept a higher Mallow's Cp, which is not ideal. However, we tend to like models with fewer variables. The final variables in this model include: speed score, line drive percentage, barrel percentage, walk percentage, strikeout percentage, and exit velocity.

With our final model in hand, we must now check and ensure our model assumptions. We can see in the residuals plots that the model passes the test. The QQ-plot shows no signs of an s-shape, and the residuals plot does not have any particular shape to it, just as we would hope.

Our final model is significant, with a p-value<0.0001. The final r-squared value is 0.6838 (68%), and the adjusted r-squared is 0.679 (67.9%). These values are shown in the table (Table:7). Every variable in the model is considered significant at the alpha=0.05 level. Our final equation that is produced is:

$$wRC + \ = \ -106.881 \ + 2.954(Spd) + 183.283(LD\%) + 1.805(EV) + 448.981(Barrel\%) + 166.706(BB\%) \ - 224.558(K\%)$$

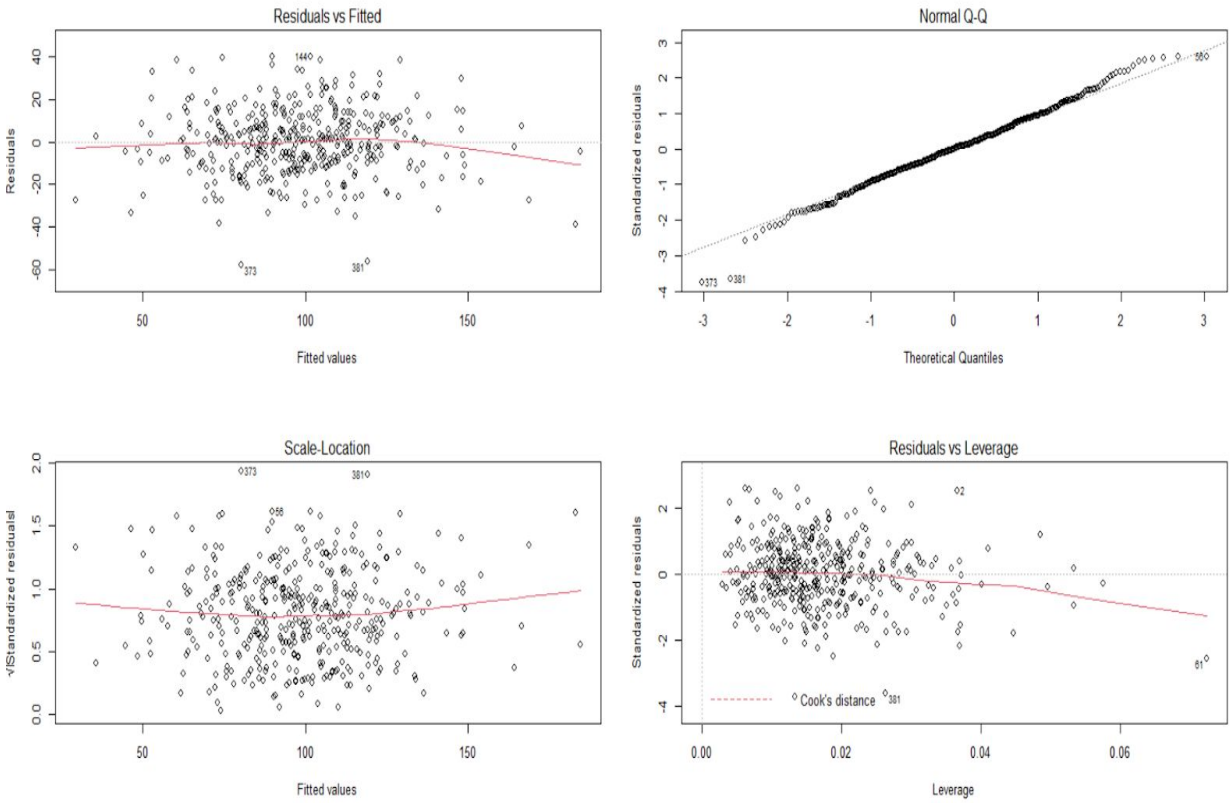


Figure 3: Residuals analysis of the model.

For each additional point of speed score, we expect a player's wRC+ to increase by 2.954. For each additional 0.01 in the proportion of batted balls that are line drives, we expect wRC+ to increase by 1.83. We expect an increase of 1.805 in wRC+ for each additional mile per hour

added to the average exit velocity. For each .01 increase in barrel percentage, we expect a bump in wRC+ of 4.49. For each increase of 0.01 in walk percentage, we expect an increase in wRC+ of 1.67. Finally, for each additional 0.01 to strikeout percentage, we expect a decrease of 2.25.

Variables	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Spd	1	1430.2677811	1430.2677811	5.8568222	0.0159627
LD%	1	18995.7361579	18995.7361579	77.7858885	< 2.2e-16
EV	1	102236.3203686	102236.3203686	418.6488456	< 2.2e-16
Barrel%	1	19656.8381521	19656.8381521	80.4930437	< 2.2e-16
BB%	1	8733.3265794	8733.3265794	35.7622132	4.956e-09
K%	1	59666.8852378	59666.8852378	244.3307088	< 2.2e-16
Residuals	399	97437.9656245	244.2054276		
Model Summary					
R^2	Adjusted R^2	P-value	F-Statistic	Residual Standard Error	
0.6838	0.679	< 2.2e-16	143.8	15.63	

Table 7: ANOVA of the updated model after dropping variables.

Model Validation

Now that we have obtained a final model and tested it for significance, it is time to test our model with future data. We will apply the model to 2020 MLB data to see how well the model's prediction of wRC+ compares to the actual outcomes of wRC+. We will generate a simple linear regression line plot using our equation for wRC+. The dependent variable that will be produced is the expected wRC+ (xwRC+) for the 2020 MLB season. As a reminder, the independent variables that will be used to find xwRC+ are speed score, line drive percentage, exit velocity, barrel percentage, walk percentage, and strikeout percentage. The sample size of

players with 100 batted ball events or more for the 2020 MLB season is 193 players. The results can be found in the fitted line plot in figure 4. As we can see, it appears that the model predicts the actual values of wRC+ for the 2020 MLB season.

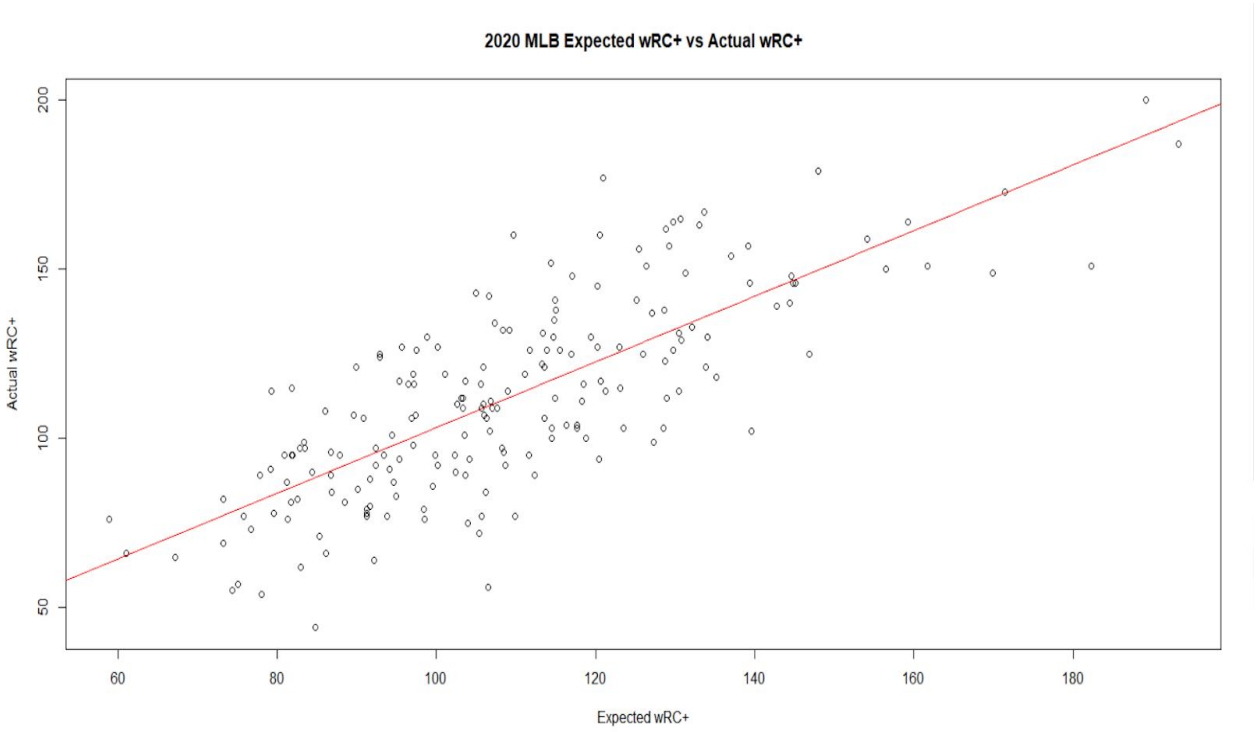


Figure 4: A fitted line plot of expected wRC+ compared to actual wRC+ values for 2020 MLB season data.

Conclusion

Baseball's statistical revolution has been a game-changer in almost every conceivable part of the game. When Statcast was released in 2015, the statistical revolution took one of it's most enormous steps yet. Statcast brought powers to baseball evolution that sabermetricians had previously dreamed of having. This paper set out to prove the statistical power that Statcast carries by explaining the variation in weighted runs created plus through the use of exit velocity, launch angle, and other Statcast variables. Through the process of linear regression, we learned that average exit velocity can significantly explain the variation in weighted runs created plus. Our model was able to explain nearly 70% of the variation in wRC+.

Average exit velocity was one of the most integral variables in the model, just as we had hypothesized. For every additional mile per hour added to average exit velocity, we expect an increase of nearly two wRC+. Considering that this is equivalent to a two percent increase in offensive production, exit velocity can be a great place to increase production. Launch angle did not have the same fate as exit velocity. The launch angle was still considered significant, but not as much as the other variables and ended up being dropped in favor of a condensed model. This does not mean that the launch angle is a useless statistic, however. This evaluation merely means that the launch angle did not add enough to the model to make it valuable to keep.

Already one of the hottest topics in baseball, Statcast research is inevitable. This paper offers some possible future issues to be investigated with this research. The most straightforward answer would be to evaluate this model in different years. As Statcast data has become more popular, the players have begun to embrace it more and more. As they have come to implement these statistics in their games, it can be assumed that their approach has changed. For example, "The average launch angle of a batted ball has increased in every year of the Statcast era, rising gradually from 10.1 degrees in 2015 to 11.7 in 2018… The average launch angle against the shift last season was 14.7 degrees, a notable jump up from 13.1 in 2015."[21] Creating a model and evaluating if these Statcast variables have significantly changed or become more significant as the years have gone by would be beneficial research for the sake of the game.

Acknowledgments

The author thanks Professor Naima Shifa for helpful comments, corrections, and references.

Sources

[1] Anonymous, Minitab.com, Enough Is Enough! Handling Multicollinearity in Regression Analysis (April 16, 2013), https://blog.minitab.com/blog/understanding-statistics/handling-multicollinearity-in-regression-analysis

[2] Anonymous, 525,600 minutes: How do you measure a player in a year? (November 14th, 2007), http://web.archive.org/web/20080102094412/http:/mvn.com/mlb-stats/2007/11/14/525600-minutes-how-do-you-measure-a-player-in-a-year/

[3] R. Arthur. FiveThirtyEight, Chase Utley Is The Unluckiest Man In Baseball (May 15, 2015) https://fivethirtyeight.com/features/chase-utley-is-the-unluckiest-man-in-baseball/

[4] E. Baccellieri, Deadspin, Major League Baseball's Statcast Can Break Sabermetrics (December 18, 2017), https://deadspin.com/major-league-baseballs-statcast-can-break-sabermetrics-1820987737

[5] R. Bevans, Scribbr, ANOVA in R: A step-by-step guide (March 6, 2020), https://www.scribbr.com/statistics/anova-in-r/

[6] J.C. Bradbury, Baseball Prospectus, How Do Baseball Players Age?: Investigating the Age-27 Theory (January 11, 2010), https://www.baseballprospectus.com/news/article/9933/how-do-baseball-players-age-investigating-the-age-27-theory/

[7] R. A. Carleton, Baseball Prospectus, Baseball Therapy: The One About Exit Velocity (April 19, 2016), https://www.baseballprospectus.com/news/article/28956/baseball-therapy-the-one-about-exit-velocity/

[8] J. Chang, J. Zenilman, A Study of Sabermetrics in Major League Baseball: The Impact of Moneyball on Free Agent Salaries, Washington University in St. Louis, 2013, https://olinblog.wustl.edu/wp-content/uploads/AStudyofSabermetricsinMajorLeagueBaseball.pdf

[9] H. Demmink III, Value of Stealing Bases in Major League Baseball: "Stealing" Runs and Wins, Public Choice , Mar., 2010, Vol. 142, No. 3/4, Essays in Honor of Robert D. Tollison (Mar., 2010), pp. 497-505, https://www.jstor.org/stable/40541986

[10] D. Fox, Fangraphs, Run Estimation for the Masses (January 12, 2006), https://tht.fangraphs.com/ops-for-the-masses/

[11] J. LeDoux, M. Schorr, GitHub, Pybaseball (October 14, 2020), https://github.com/jldbc/pybaseball

[12] MLB.com, “Exit Velocity (EV)” http://m.mlb.com/glossary/statcast/exit-velocity

[13] MLB.com, “Launch Angle (LA)” http://m.mlb.com/glossary/statcast/launch-angle

[14] MLB.com, “Statcast|Glossary,” http://m.mlb.com/glossary/statcast

[15] MLB.com, “Barrel” http://m.mlb.com/glossary/statcast/barrel

[16] D. C. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis.* John Wiley, 2012

[17] B. Petti, Fangraphs, Using Statcast Data to Predict Hits (June 14, 2016), https://tht.fangraphs.com/using-statcast-data-to-predict-hits/

[18] C. J. Phillips, Harvard Data Science Review, The Bases of Data (November 01, 2019). https://hdsr.mitpress.mit.edu/pub/3adoxb26/release/2?readingCollection=af83430a

[19] D. Richards, Pitchers List, Going Deep: The Real Value of Statcast Data Part I (2018), https://www.pitcherlist.com/going-deep-the-real-value-of-statcast-data-part-i/

[20] G. D. Sanford, What raw statistics have the greatest effect on wRC+ in Major League Baseball in 2017?, https://conservancy.umn.edu/bitstream/handle/11299/199934/Sanford%2C%20Gavin%20%28What%20raw%20statistics%20have%20the%20greatest%20effect%20on%20wRC%2B%20in%20Major%20League%20Baseball%20in%202017%29Capstone%20201718.pdf?sequence=1&isAllowed=y

[21] T. Sawchik, FiveThirtyEight, Don’t Worry, MLB — Hitters Are Killing The Shift On Their Own (January 17, 2019), https://fivethirtyeight.com/features/dont-worry-mlb-hitters-are-killing-the-shift-on-their-own/#:~:text=Consider%20that%20in%202018%2C%2032.7,the%20pull%20side%20and%20a%20.

[22] D. Sheinin, A. Emamdjomeh, Washington Post, These days in baseball, every batter is trying to find an angle (June 1, 2017), https://www.washingtonpost.com/graphics/sports/mlb-launch-angles-story/#:~:text=Analysts%20have%20been%20able%20to,of%2095%20mph%20or%20greater

[23] S. Slowinski, Fangraphs, GB%, LD%, FB% (February 17, 2010), https://library.fangraphs.com/pitching/batted-ball/

[24] S. Slowinski, Fangraphs, Spd (February 16, 2010) https://library.fangraphs.com/offense/spd/

[25] S. Slowinski, Fangraphs, wRC and wRC+ (February 16, 2010), https://library.fangraphs.com/offense/wrc/

[26] B. Stampfl, Fangraphs, Barrels, Normative Analysis, and the Beauties of Statcast (September 29, 2016), https://tht.fangraphs.com/barrels-normative-analysis-and-the-beauties-of-statcast/

[27] N. C. Taylor, Forecasting Batter Performance Using Statcast Data In Major League Baseball, (2017), https://library.ndsu.edu/ir/bitstream/handle/10365/28371/Taylor_ndsu_0157N_11679.pdf?sequence=1&isAllowed=y

[28] J. Trupin, SB Nation, An Idiot's Guide to Advanced Statistics: wOBA and wRC+ (Mar 7, 2017), https://www.lookoutlanding.com/2017/3/7/14783982/an-idiots-guide-to-advanced-statistics-woba-and-wrc-sabermetrics

[29] T. L. Turocy, An inspection game model of the stolen base in baseball: A theory of theft, University of East Anglia, 2014, http://www.gambit-project.org/turocy/papers/theft-20140822.pdf

[30] N. Weinberg, Fangraphs, Complete List (Offense) (October 30, 2014), https://library.fangraphs.com/offense/offensive-statistics-list/

[31] W. L. Winston, *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*, Princeton University Press, 2020, P. 17-29 https://www-jstor-org.duproxy.palni.edu/stable/pdf/j.ctt7sj9q.8.pdf?ab_segments=0%2Fbasic_search_solr_cloud%2Fcontrol&refreqid=fastly-default%3Ac1cee0ae3c654702ceaf80c2da7d04a3

Appendix

Python

```
pip install pybaseball
from pybaseball import statcast_batter_exitvelo_barrels, batting_stats
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns
import warnings
import feather
warnings.filterwarnings('ignore')
%matplotlib inline

data = batting_stats(2019, qual=0)
data.head()
data.count()
```

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
print(data.columns.tolist())

df = pd.DataFrame(data)
column = 'Events'
df2 = df[df[column].between(100, 100000)]
df2.describe()

final_data = df2[['wRC+', 'Age','BB%','K%','Spd', 'Pull%', 'Cent%',
'Oppo%', 'LD%','GB%','FB%','EV','LA','Barrel%']]

compression_opts = dict(method='zip',archive_name='out.csv')
final_data.to_csv('out.zip', index=False,compression=compression_opts)

data = batting_stats(2020, qual=0)
data.head()
data.count()

df20 = pd.DataFrame(data)
column = 'Events'
df20_2 = df20[df20[column].between(100, 1000000)]
df20_2.describe()

final_data_2020 = df202[['wRC+','BB%','K%','Spd', 'LD%','EV','Barrel%']]

compression_opts = dict(method='zip',archive_name='out.csv')
final_data_2020.to_csv('out.zip',index=False,compression=compression_opts)
```

R-Code

```
##Input Data
setwd("C:/Users/Owner/Desktop/2020 Classes/Seminar/Project 2")
data <- read.csv("2019_wRC+.csv")
data_2020 <- read.csv("2020_xwRC+.csv")

##Install Necessary Packages
install.packages("car")
install.packages("olsrr")
install.packages("MPV")
install.packages("leaps")
install.packages("dplyr")
install.packages("ggplot2")

##Install Necessary Libraries
library("ggplot2")
library("dplyr")
library(olsrr)
```

```
library(MASS)
library(MPV)
library(leaps)

##DGP
wRC+ = data$wRC+
wRC._2020 = data_2020$wRC+

data
dataset_2 = data_2020[,2:7]


search()
attach(data)
search()

search()
attach(data_2020)
search()


##Correlation matrix
var <- c("wRC.", "Age", "Spd", "Pull.", "Oppo.", "LD.", "FB.", "EV", "LA", "Barrel.", "BB.",
"K.")
depVars <- data[var]
res <- cor(depVars)
round(res, 3)
write.table(round(res,3), file="corMatrix1.txt", row.names=TRUE, col.names=TRUE, sep="\t")


##Have to check the normality assumption of the dependent variable to run anova
hist(wRC.)
qqnorm(wRC., pch = 1, frame = FALSE, main="wRC+ QQ-plot")
qqline(wRC., col = "steelblue", lwd = 2)


##Initial Model
mod1 <- lm(wRC. ~ Age + Spd + Pull. + Oppo. + LD. + FB. + EV + LA + Barrel. + BB. + K.)


write.table(round(anova(mod1),7), file="anovaMod1.txt", row.names=TRUE, col.names=TRUE,
sep="\t")
anova(mod1)
summary(mod1)

##VIFs
```

```
write.table(round(car::vif(mod1),4), file="VIFMod1.txt", row.names=TRUE, col.names=TRUE,
sep="\t")

##New model
mod2 <- lm(wRC. ~ Age + Spd + Pull. + Oppo. + LD. + LA + EV + Barrel. + BB. + K.)
write.table(round(car::vif(mod2),4), file="VIFMod3.txt", row.names=TRUE, col.names=TRUE,
sep="\t")

##Stepwise Regression
ols_step_all_possible(mod2, details = TRUE)

##Forward Regression
ols_step_forward_p(mod2, details = TRUE)
write.table(as.data.frame(ols_step_forward_p(mod2)), file="SForMod3.txt", row.names=TRUE,
col.names=TRUE, sep="\t")

##Backward Regression
step(mod2, direction = "backward", details=TRUE )
step.model <- stepAIC(mod2, direction = "backward", trace = FALSE)
summary(step.model)

##Final Model
fin_mod <- lm(wRC. ~ Spd + LD. + EV + Barrel. + BB. + K.)
par(mfrow = c(2, 2))
plot(fin_mod)

par(mfrow = c(2, 2))
plot(wRC.)

anova(fin_mod)

summary(fin_mod)

write.table(round(anova(fin_mod),7), file="Fin_Mod.txt", row.names=TRUE, col.names=TRUE,
sep="\t")

##Model Validation

model.frame(wRC. ~ Spd + LD. + EV + Barrel. + BB. + K., data = data_2020)

wRC_2 <- data_2020$wRC
xwRC <- predict(fin_mod, newdata = data_2020)

preMod <- lm(wRC_2 ~ xwRC)

summary(preMod)
```

```
plot(xwRC, wRC_2, main="2020 MLB Expected wRC+ vs Actual wRC+", xlab="Expected
wRC+", ylab="Actual wRC+")
abline(preMod, col="red")
```